Discovering Discrete Latent Topics with Neural Variational Inference

Yishu Miao [1] Edward Grefenstette [2] Phil Blunsom [1 2]

Abstract

Topic models have been widely explored as probabilistic generative models of documents. Traditional inference methods have sought closed-form derivations for updating the models, however as the expressiveness of these models grows, so does the difficulty of performing fast and accurate inference over their parameters. This paper presents alternative neural approaches to topic modelling by providing parameterisable distributions over topics which permit training by backpropagation in the framework of neural variational inference. In addition, with the help of a stick-breaking construction, we propose a recurrent network that is able to discover a notionally unbounded number of topics, analogous to Bayesian non-parametric topic models. Experimental results on the MXM Song Lyrics, 20NewsGroups and Reuters News datasets demonstrate the effectiveness and efficiency of these neural topic models.

1. Introduction

Probabilistic models for inducing latent topics from documents are one of the great success stories of unsupervised learning. Starting with latent semantic analysis (LSA (Landauer et al., 1998)), models for uncovering the underlying semantic structure of a document collection have been widely applied in data mining, text processing and information retrieval. Probabilistic topic models (e.g. PLSA (Hofmann, 1999), LDA (Blei et al., 2003) and HDPs (Teh et al., 2006)) provide a robust, scalable, and theoretically sound foundation for document modelling by introducing latent variables for each token to topic assignment.

For the traditional Dirichlet-Multinomial topic model, efficient inference is available by exploiting conjugacy with either Monte Carlo or Variational techniques (Jordan et al., 1999; Attias, 2000; Beal, 2003)). However, as topic models have grown more expressive, in order to capture topic dependencies or exploit conditional information, inference methods have become increasingly complex. This is especially apparent for non-conjugate models (Carlin & Polson, 1991; Blei & Lafferty, 2007; Wang & Blei, 2013).

Deep neural networks are excellent function approximators and have shown great potential for learning complicated non-linear distributions for unsupervised models. Neural variational inference (Kingma & Welling, 2014; Rezende et al., 2014; Mnih & Gregor, 2014) approximates the posterior of a generative model with a variational distribution parameterised by a neural network. This allows both the generative model and the variational network to be jointly trained with backpropagation. For models with continuous latent variables associated with particular distributions, such as Gaussians, there exist reparameterisations (Kingma & Welling, 2014; Rezende et al., 2014) of the distribution permitting unbiased and low-variance estimates of the gradients with respect to the parameters of the inference network. For models with discrete latent variables, Monte-Carlo estimates of the gradient must be employed. Recently, algorithms such as REINFORCE have been used effectively to decrease variance and improve learning (Mnih & Gregor, 2014; Mnih et al., 2014).

In this work we propose and evaluate a range of topic models parameterised with neural networks and trained with variational inference. We introduce three different neural structures for constructing topic distributions: the Gaussian Softmax distribution (GSM), the Gaussian Stick Breaking distribution (GSB), and the Recurrent Stick Breaking process (RSB), all of which are conditioned on a draw from a multivariate Gaussian distribution. The Gaussian Softmax topic model constructs a finite topic distribution with a softmax function applied to the projection of the Gaussian random vector. The Gaussian Stick Breaking model also constructs a discrete distribution from the Gaussian draw, but this time employing a stick breaking construction to provide a bias towards sparse topic distributions. Finally, the Recurrent Stick Breaking process employs a recurrent neural network, again conditioned on the Gaussian draw, to progressively break the stick, yielding a neural analog of a Dirichlet Process topic model (Teh et al., 2006).

[1] University of Oxford, Oxford, United Kingdom [2] DeepMind, London, United Kingdom. Correspondence to: Yishu Miao <yishu.miao@cs.ox.ac.uk>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP. Copyright 2017 by the author(s).

Our neural topic models combine the merits of both neural networks and traditional probabilistic topic models. They can be trained efficiently by backpropagation, scaled to large data sets, and easily conditioned on any available contextual information. Further, as probabilistic graphical models, they are interpretable and explicitly represent the dependencies amongst the random variables. Previous neural document models, such as the neural variational document model (NVDM) (Miao et al., 2016), belief networks document model (Mnih & Gregor, 2014), neural auto-regressive document model (Larochelle & Lauly, 2012) and replicated softmax (Hinton & Salakhutdinov, 2009), have not explicitly modelled latent topics. Through evaluations on a range of data sets we compare our models with previously proposed neural document models and traditional probabilistic topic models, demonstrating their robustness and effectiveness.

2. Parameterising Topic Distributions

In probabilistic topic models, such as LDA (Blei et al., 2003), we use the latent variables θ_d and z_n for the topic proportion of document d, and the topic assignment for the observed word w_n, respectively. In order to facilitate efficient inference, the Dirichlet distribution (or Dirichlet process (Teh et al., 2006)) is employed as the prior to generate the parameters of the multinomial distribution θ_d for each document. The use of a conjugate prior allows the tractable computation of the posterior distribution over the latent variables' values. While alternatives have been explored, such as log-normal topic distributions (Blei & Lafferty, 2006; 2007), extra approximation (e.g. the Laplace approximation (Wang & Blei, 2013)) is required for closed form derivations. The generative process of LDA is:

$$\begin{aligned} \theta_d &\sim \text{Dir}(\alpha_0), && \text{for } d \in D \\ z_n &\sim \text{Multi}(\theta_d), && \text{for } n \in [1, N_d] \\ w_n &\sim \text{Multi}(\beta_{z_n}), && \text{for } n \in [1, N_d] \end{aligned}$$

where β_{z_n} represents the topic distribution over words given topic assignment z_n and N_d is the number of tokens in document d. β_{z_n} can be drawn from another Dirichlet distribution, but here we consider it a model parameter. α_0 is the hyper-parameter of the Dirichlet prior and N_d is the total number of words in document d. The marginal likelihood for a document in collection D is:

$$p(d|\alpha_0, \beta) = \int_\theta p(\theta|\alpha_0) \prod_n \sum_{z_n} p(w_n|\beta_{z_n}) p(z_n|\theta) d\theta. \qquad (1)$$

If we employ mean-field variational inference, the updates for the variational parameters $q(\theta)$ and $q(z_n)$ can be directly derived in closed form.

In contrast, our proposed models introduce a neural network to parameterise the multinomial topic distribution. The generative process is:

$$\begin{aligned} \theta_d &\sim \text{G}(\mu_0, \sigma_0^2), && \text{for } d \in D \\ z_n &\sim \text{Multi}(\theta_d), && \text{for } n \in [1, N_d] \\ w_n &\sim \text{Multi}(\beta_{z_n}), && \text{for } n \in [1, N_d] \end{aligned}$$

where $G(\mu_0, \sigma_0^2)$ is composed of a neural network $\theta = g(x)$ conditioned on a isotropic Gaussian $x \sim N(\mu_0, \sigma_0^2)$[1]. The marginal likelihood is:

$$\begin{aligned} p(d|\mu_0, \sigma_0, \beta) \;=\; & \int_\theta p(\theta|\mu_0, \sigma_0^2) \qquad (2) \\ & \prod\nolimits_n \sum\nolimits_{z_n} p(w_n|\beta_{z_n}) p(z_n|\theta) d\theta. \end{aligned}$$

Compared to Equation (1), here we parameterise the latent variable θ by a neural network conditioned on a draw from a Gaussian distribution. To carry out neural variational inference (Miao et al., 2016), we construct an inference network $q(\theta|\mu(d), \sigma(d))$ to approximate the posterior $p(\theta|d)$, where $\mu(d)$ and $\sigma(d)$ are functions of d that are implemented by multilayer perceptrons (MLP). By using a Gaussian prior distribution, we are able to employ the re-parameterisation trick (Kingma & Welling, 2014) to build an unbiased and low-variance gradient estimator for the variational distribution. Without conjugacy, the updates of the parameters can still be derived directly and easily from the variational lower bound. We defer discussion of the inference process until the next section. Here we introduce several alternative neural networks for $g(x)$ which transform a Gaussian sample x into the topic proportions θ.

2.1. The Gaussian Softmax Construction

In deep learning, an energy-based function is generally used to construct probability distributions (LeCun et al., 2006). Here we pass a Gaussian random vector through a softmax function to parameterise the multinomial document topic distributions. Thus $\theta \sim G_{\text{GSM}}(\mu_0, \sigma_0^2)$ is defined as:

$$\begin{aligned} x &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \theta &= \text{softmax}(W_1^T x) \end{aligned}$$

where W_1 is a linear transformation, and we leave out the bias terms for brevity. μ_0 and σ_0^2 are hyper-parameters which we set for a zero mean and unit variance Gaussian.

2.2. The Gaussian Stick Breaking Construction

In Bayesian non-parametrics, the stick breaking process (Sethuraman, 1994) is used as a constructive definition of the Dirichlet process, where sequentially drawn Beta

[1] Throughout this presentation we employ diagonal Gaussian distributions. As such we use $N(\mu, \sigma^2)$ to represent the Gaussian distributions, where σ^2 is the diagonal of the covariance matrix.

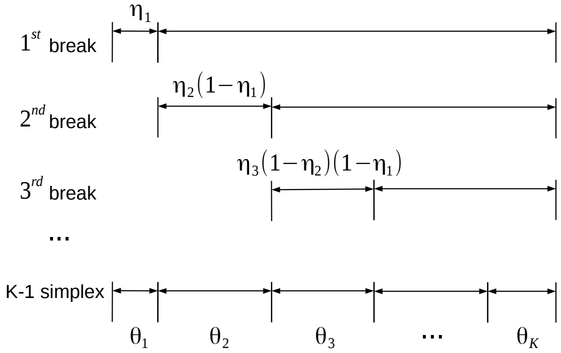


Figure 1. The Stick Breaking Construction.

random variables define breaks from a unit stick. In our case, following Khan et al. (2012), we transform the modelling of multinomial probability parameters into the modelling of the logits of binomial probability parameters using Gaussian latent variables. More specifically, conditioned on a Gaussian sample $x \in \mathbb{R}^H$, the breaking proportions $\eta \in \mathbb{R}^{K-1}$ are generated by applying the sigmoid function $\eta = \text{sigmoid}(W_2^T x)$ where $W \in \mathbb{R}^{H \times K-1}$. Starting with the first piece of the stick, the probability of the first category is modelled as a break of proportion η_1, while the length of the remainder of the stick is left for the next break. Thus each dimension can be deterministically computed by $\theta_k = \eta_k \prod_{i=1}^{k-1}(1-\eta_i)$ until $k = K-1$, and the remaining length is taken as the probability of the Kth category $\theta_K = \prod_{i=1}^{K-1}(1-\eta_i)$.

For instance assume $K = 3$, θ is generated by 2 breaks where $\theta_1 = \eta_1$, $\theta_2 = \eta_2(1-\eta_1)$ and the remaining stick $\theta_3 = (1-\eta_2)(1-\eta_1)$. If the model proceeds to break the stick for $K = 4$, the remaining stick θ_3 is broken into (θ'_3, θ'_4), where $\theta'_3 = \eta_3 \cdot \theta_3$, $\theta'_4 = (1-\eta_3) \cdot \theta_3$ and $\theta_3 = \theta'_3 + \theta'_4$. Hence, for different values of K, it always satisfies $\sum_{k=1}^{K} \theta_k = 1$. The stick breaking construction $f_{\text{SB}}(\eta)$ is illustrated in Figure 1 and the distribution $\theta \sim G_{\text{GSB}}(\mu_0, \sigma_0^2)$ is defined as:

$$\begin{aligned} x &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \eta &= \text{sigmoid}(W_2^T x) \\ \theta &= f_{\text{SB}}(\eta) \end{aligned}$$

Although the Gaussian stick breaking construction breaks exchangeability, compared to the stick breaking definition of the Dirichlet process, it does provide a more amenable form for neural variational inference. More interestingly, this stick breaking construction introduces a non-parametric aspect to neural topic models.

2.3. The Recurrent Stick Breaking Construction

Recurrent Neural Networks (RNN) are commonly used for modelling sequences of inputs in deep learning. Here we consider the stick breaking construction as a sequential draw from an RNN, thus capturing an unbounded number of breaks with a finite number of parameters. Conditioned on a Gaussian latent variable x, the recurrent neural network $f_{\text{SB}}(x)$ produces a sequence of binomial logits which are used to break the stick sequentially. The $f_{\text{RNN}}(x)$ is decomposed as:

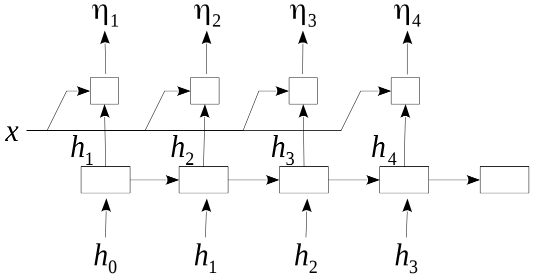


Figure 2. The unrolled Recurrent Neural Network that produces the stick breaking proportions η.

$$\begin{aligned} h_k &= \text{RNN}_{\text{SB}}(h_{k-1}) \\ \eta_k &= \text{sigmoid}(h_{k-1}^T x) \end{aligned}$$

where h_k is the output of the kth state, which we feed into the next state of the RNN_{SB} as an input. Figure 2 shows the recurrent neural network structure. Now $\theta \sim G_{\text{RSB}}(\mu_0, \sigma_0^2)$ is defined as:

$$\begin{aligned} x &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \eta &= f_{\text{RNN}}(x) \\ \theta &= f_{\text{SB}}(\eta) \end{aligned}$$

where $f_{\text{SB}}(\eta)$ is equivalent to the stick breaking function used in the Gaussian stick breaking construction. Here, the RNN is able to dynamically produce new logits to break the stick *ad infinitum*. The expressive power of the RNN to model sequences of unbounded length is still bounded by the parametric model's capacity, but for topic modelling it is adequate to model the countably infinite topics amongst the documents in a truncation-free fashion.

3. Models

Given the above described constructions for the topic distributions, in this section we introduce our family of neural topic models and corresponding inference methods.

3.1. Neural Topic Models

Assume we have finite number of topics K, the topic distribution over words given a topic assignment z_n is $p(w_n|\beta, z_n) = \text{Multi}(\beta_{z_n})$. Here we introduce topic vectors $t \in \mathbb{R}^{K \times H}$, word vectors $v \in \mathbb{R}^{V \times H}$ and generate the topic distributions over words by:

$$\beta_k = \text{softmax}(v \cdot t_k^T).$$

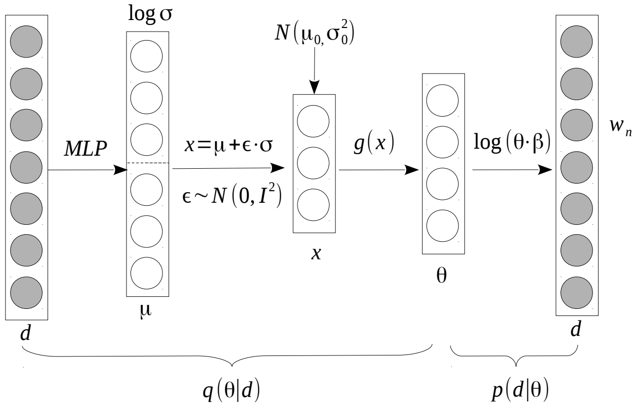


Figure 3. Network structure of the inference model $q(\theta \mid d)$, and of the generative model $p(d \mid \theta)$.

Therefore, $\beta \in \mathbb{R}^{K\times V}$ is a collection of simplexes achieved by computing the semantic similarity between topics and words. Following the notation introduced in Section 2, the prior distribution is defined as $G(\theta|\mu_0, \sigma_0^2)$ in which $x \sim \mathcal{N}(x|\mu_0, \sigma_0^2)$ and the projection network generates $\theta = g(x)$ for each document. Here, $g(x)$ can be the Gaussian Softmax $g_{\text{GSM}}(x)$, Gaussian Stick Breaking $g_{\text{GSB}}(x)$, or Recurrent Stick Breaking $g_{\text{RSB}}(x)$ constructions with fixed length RNN_{SB}. We derive a variational lower bound for the document log-likelihood according to Equation (2):

$$\begin{aligned}\mathcal{L}_d = \ & \mathbb{E}_{q(\theta|d)}\left[\sum\nolimits_{n=1}^{N} \log \sum\nolimits_{z_n} [p(w_n|\beta_{z_n})p(z_n|\theta)]\right] \\ & -D_{KL}\left[q(\theta|d)||p(\theta|\mu_0,\sigma_0^2)\right] \qquad (3)\end{aligned}$$

where $q(\theta|d)$ is the variational distribution approximating the true posterior $p(\theta|d)$. Following the framework of neural variational inference (Miao et al., 2016; Kingma & Welling, 2014; Rezende et al., 2014), we introduce an inference network conditioned on the observed document d to generate the variational parameters $\mu(d)$ and $\sigma(d)$ so that we can estimate the lower bound by sampling θ from $q(\theta|d) = G(\theta|\mu(d), \sigma^2(d))$. In practise we reparameterise $\hat{\theta} = \mu(d) + \hat{\epsilon}\cdot\sigma(d)$ with the sample $\hat{\epsilon} \in \mathcal{N}(0, I)$.

Since the generative distribution $p(\theta|\mu_0, \sigma_0^2) = p(g(x)|\mu_0, \sigma_0^2) = p(x|\mu_0, \sigma_0^2)$ and the variational distribution $q(\theta|d) = q(g(x)|d) = q(x|\mu(d), \sigma^2(d))$, the KL term in Equation (3) can be easily integrated as a Gaussian KL-divergence. Note that, the parameterisation network $g(x)$ and its parameters are shared across all the documents. In addition, given a sampled $\hat{\theta}$, the latent variable z_n can be integrated out as:

$$\begin{aligned}\log p(w_n|\beta,\hat{\theta}) &= \log \sum\nolimits_{z_n}\left[p(w_n|\beta_{z_n})p(z_n|\hat{\theta})\right] \\ &= \log(\hat{\theta}\cdot\beta) \qquad (4)\end{aligned}$$

Thus there is no need to introduce another variational approximation for the topic assignment z. The variational lower bound is therefore:

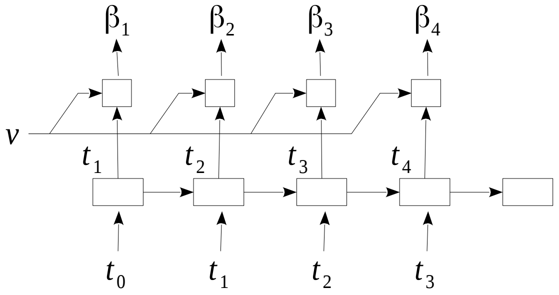


Figure 4. The unrolled Recurrent Neural Network that produces the topic-word distributions β.

$$\mathcal{L}_d \approx \hat{\mathcal{L}}_d = \sum\nolimits_{n=1}^{N}\left[\log p(w_n|\beta,\hat{\theta})\right] - D_{KL}\left[q(x|d)||p(x)\right]$$

We can directly derive the gradients of the generative parameters Θ, including t, v and $g(x)$. While for the variational parameters Φ, including $\mu(d)$ and $\sigma(d)$, we use the gradient estimators:

$$\begin{aligned}\nabla_{\mu(d)}\hat{\mathcal{L}}_d &\approx \nabla_{\hat{\theta}}\hat{\mathcal{L}}_d, \\ \nabla_{\sigma(d)}\hat{\mathcal{L}}_d &\approx \hat{\epsilon}\cdot\nabla_{\hat{\theta}}\hat{\mathcal{L}}_d.\end{aligned}$$

Θ and Φ are jointly updated by stochastic gradient backpropagation. The structure of this variational auto-encoder is illustrated in Figure 3.

3.2. Recurrent Neural Topic Models

For the GSM and GSB models the topic vectors $t \in \mathbb{R}^{K\times H}$ have to be predefined for computing the topic distribution over words β. With the RSB construction we can model an unbounded number of topics, however in addition to the RNN_{SB} that generates the topic proportions $\theta \in \mathbb{R}^{\infty}$ for each document, we must introduce another neural network $\text{RNN}_{\text{Topic}}$ to produce the topics $t \in \mathbb{R}^{\infty\times H}$ dynamically, so as to avoid the need to truncate the variational inference.

For comparison, in finite neural topic models we have topic vectors $t \in \mathbb{R}^{K\times H}$, while in unbounded neural topic models the topics $t \in \mathbb{R}^{\infty\times H}$ are dynamically generated by $\text{RNN}_{\text{Topic}}$ and the order of the topics corresponds to the order of the states in RNN_{SB}. The generation of β follows:

$$\begin{aligned}t_k &= \text{RNN}_{\text{Topic}}(t_{k-1}), \\ \beta_k &= \text{softmax}(v\cdot t_k^T),\end{aligned}$$

where $v \in \mathbb{R}^{V\times H}$ represents the word vectors, t_k is the kth topic generated by $\text{RNN}_{\text{Topic}}$ and $k < \infty$. Figure 4 illustrates the neural structure of $\text{RNN}_{\text{Topic}}$.

For the unbounded topic models we introduce a truncation-free neural variational inference method which enables the model to dynamically decide the number of active topics. Assume the current active number of topics is i, $\text{RNN}_{\text{Topic}}$ generates $t^i \in \mathbb{R}^{i \times H}$ by an $i-1$ step stick-breaking process (the logit for the nth topic is the remaining stick after $i-1$ breaks). The variational lower bound for a document d is:

$$\mathcal{L}_d^i \approx \sum\nolimits_{n=1}^{N} \left[\log p(w_n | \beta^i, \hat{\theta}^i) \right] - D_{KL} \left[q(x|d) || p(x) \right],$$

where $\hat{\theta}^i$ corresponds to the topic distribution over words β^t. In order to dynamically increase the number of topics, the model proposes the ith break on the stick to split the $(i+1)$th topic. In this case, $\text{RNN}_{\text{Topic}}$ proceeds to the next state and generates topic $t^{(i+1)}$ for $\beta^{(i+1)}$ and the RNN_{SB} generates $\hat{\theta}^{(i+1)}$ by an extra break of the stick. Firstly, we compute the likelihood increase brought by topic i across the documents D:

$$\mathcal{I} = \sum\nolimits_{d}^{D} \left[\mathcal{L}_d^i - \mathcal{L}_d^{i-1} \right] / \sum\nolimits_{d}^{D} [\mathcal{L}_d^i]$$

Then, we employ an acceptance hyper-parameter γ to decide whether to generate a new topic. If $\mathcal{I} > \gamma$, the previous proposed new topic (the ith topic) contributes to the generation of words and we increase the active number of topics i by 1, otherwise we keep the current i unchanged. Thus γ controls the rate at which the model generates new topics. In practise, the increase of the lower bound is computed over mini-batches so that the model is able to generate new topics before the current epoch is finished. The details of the algorithm are described in Algorithm 1.

3.3. Topic vs. Document Models

In most topic models, documents are modelled by a mixture of topics, and each word is associated with a single topic latent variable, e.g. LDA and GSM. However, the NVDM (Miao et al., 2016) is implemented as a VAE (Kingma & Welling, 2014) without modelling topics explicitly. The major difference is that NVDM employs a softmax decoder (Equation 5) to generate all of the words of a document conditioned on the document representation $\hat{\theta}$:

$$\log p(w_n | \beta, \hat{\theta}) = \log \text{softmax}(\hat{\theta} \cdot \beta). \tag{5}$$

where both $\hat{\theta}$ and β are unnormalised. Hence, it breaks the topic model assumption that each document consists of a mixture of topics. Although the latent variables can still be interpreted as topics, these topics are not modelled explicitly since there is no actual topic distribution over words. Srivastava & Sutton (2016) interprets the above decoder as a weighted product of experts topic model, here however, we refer to such models that do not directly assign topics to words as document models instead of topic models. We can also convert our neural topic models to neural document models by replacing the mixture decoder (Equation 4) with the softmax decoder (Equation 5). For example in the GSM construction, if we remove the softmax function over θ, and directly apply Equation 5 to generate the words, it reduces to a variant of the NVDM (GSM applies topic and word vectors to compute β, while NVDM directly models a projection β from the latent variables to words).

Algorithm 1 Unbounded Recurrent Neural Topic Model

```
Initialise $\Theta$ and $\Phi$; Set active topic number $i$
repeat
  for $s \in$ minibatches $S$ do
    for $k \in [1, i]$ do
      Compute topic vector $t_k = \text{RNN}_{\text{Topic}}(t_{k-1})$
      Compute topic distribution $\beta_k = \text{softmax}(v \cdot t_k^T)$
    end for
    for $d \in D_s$ do
      Sample topic proportion $\hat{\theta} \sim G_{\text{RSB}}(\theta | \mu(d), \sigma^2(d))$
      for $w \in$ document $d$ do
        Compute log-likelihood $\log p(w | \hat{\theta}, \beta)$
      end for
      Compute lowerbound $\mathcal{L}_d^{i-1}$ and $\mathcal{L}_d^i$
      Compute gradients $\nabla_{\Theta,\Phi} \mathcal{L}_d^i$ and update
    end for
    Compute likelihood increase $\mathcal{I}$
    if $\mathcal{I} > \gamma$ then
      Increase active topic number $i = i + 1$
    end if
  end for
until Convergence
```

4. Related Work

Topic models have been extensively studied for a variety of applications in document modelling and information retrieval. Beyond LDA, significant extensions have sought to capture topic correlations (Blei & Lafferty, 2007), model temporal dependencies (Blei & Lafferty, 2006) and discover an unbounded number of topics (Teh et al., 2006). Topic models have been extended to capture extra context information such as time (Wang & McCallum, 2006), authorship (Rosen-Zvi et al., 2004), and class labels (Mcauliffe & Blei, 2008). Such extensions often require carefully tailored graphical models, and associated inference algorithms, to capture the desired context. Neural models provide a more generic and extendable option and a number of works have sought to leverage these, such as the Replicated Softmax (Hinton & Salakhutdinov, 2009), the Auto-Regressive Document Model (Larochelle & Lauly, 2012), Sigmoid Belief Document Model (Mnih & Gregor, 2014), Variational Auto-Encoder Document Model (NVDM) (Miao et al., 2016) and TopicRNN Model (Dieng et al., 2016). However, these neural works do not explicitly capture topic assignments.

The recent work of Srivastava & Sutton (2016) also employs neural variational inference to train topic models and is closely related to our work. Their model follows the original LDA formulation in keeping the Dirichlet-Multinomial

Finite Topic Model	MXM		20News		RCV1	
	50	200	50	200	50	200
GSM	**306**	**272**	**822**	830	**717**	**602**
GSB	309	296	838	826	788	634
RSB	311	297	835	**822**	750	628
OnlineLDA (Hoffman et al., 2010)	312	342	893	1015	1062	1058
NVLDA (Srivastava & Sutton, 2016)	330	357	1073	993	791	797
Unbounded Topic Model	MXM		20News		RCV1	
RSB-TF	**303**		**825**		**622**	
HDP (Wang et al., 2011)	370		937		918	

Table 1. Perplexities of the topic models on the test datasets. The upper section of the table lists the results for finite neural topic models, with 50 or 200 topics, on the MXM, 20NewsGroups and RCV1 datasets. We compare our neural topic models with the Gaussian Softmax (GSM), Gaussian Stick Breaking (GSB) and Recurrent Stick Breaking (RSB) constructions to the online variational LDA (onlineLDA) (Hoffman et al., 2010) and neural variational inference LDA (NVLDA) (Srivastava & Sutton, 2016) models. The lower section shows the results for the unbounded topic models, including our truncation-free RSB (RSB-TF) and the online HDP topic model (Wang et al., 2011).

parameterisation and applies a Laplace approximation to allow gradient to back-propagate to the variational distribution. In contrast, our models directly parameterise the multinomial distribution with neural networks and jointly learn the model and variational parameters during inference. Nalisnick & Smyth (2016) proposes a reparameterisation approach for continuous latent variables with Beta prior, which enables neural variational inference for Dirichlet process. However, Taylor expansion is required to approximate the KL Divergence while having multiple draws from the Kumaraswamy variational distribution. In our case, we can easily apply the Gaussian reparametersation trick with only one draw from the Gaussian distribution.

5. Experiments

We perform an experimental evaluation employing three datasets: *MXM*[2] song lyrics, *20NewsGroups*[3] and Reuters *RCV1-v2*[4] news. *MXM* is the official lyrics collection of the Million Song Dataset with 210,519 training and 27,143 testing datapoints respectively. The *20NewsGroups* corpus is divided into 11,314 training and 7,531 testing documents, while the *RCV1-v2* corpus is a larger collection with 794,414 training and 10,000 test cases from Reuters newswire stories. We employ the original 5,000 vocabulary provided for *MXM*, while the other two datasets are processed by stemming, filtering stopwords and taking the most frequent 2,000[5] and 10,000 words as the vocabularies.

Finite Document Model	MXM		20News		RCV1	
	50	200	50	200	50	200
GSM	**270**	**267**	787	829	**653**	**521**
GSB	285	275	816	815	712	544
RSB	286	283	**785**	**792**	662	534
NVDM (Miao et al., 2016)	345	345	837	873	717	588
ProdLDA (Srivastava & Sutton, 2016)	319	326	1009	989	780	788
Unbounded Document Model	MXM		20News		RCV1	
RSB-TF	285		788		532	

Table 2. Perplexities of document models on the test datasets. The table compares the results for a fixed dimension latent variable, 50 or 200, achieved by our neural document models to Product of Experts LDA (prodLDA) (Srivastava & Sutton, 2016) and the Neural Variational Document Model (NVDM) (Miao et al., 2016).

The hidden dimension of the MLP for constructing $q(\theta|d)$ is 256 for all the neural topic models and the benchmarks that apply neural variational inference (e.g. NVDM, proLDA, NVLDA), and 0.8 dropout is applied on the output of the MLP before parameterising the diagonal Gaussian distribution. Grid search is carried out on learning rate and batch size for achieving the held-out perplexity. For the recurrent stick breaking construction we use a one layer LSTM cell (256 hidden units) for constructing the recurrent neural network. For the finite topic models we set the maximum number of topics K as 50 and 200. The models are trained by Adam (Kingma & Ba, 2015) and only one sample is used for neural variational inference. We follow the optimisation strategy of Miao et al. (2016) by alternately updating the model parameters and the inference network. To alleviate the redundant topics issue, we also apply topic diversity regularisation (Xie et al., 2015) while carrying out neural variational inference (Appendix B).

5.1. Evaluation

We use Perplexity as the main metric for assessing the generalisation ability of our generative models. Here we use the variational lower bound to estimate the document perplexity: $\exp(-\frac{1}{D}\sum_d^D \frac{1}{N_d}\log p(d))$ following Miao et al. (2016). Table 1 presents the test document perplexities of the topic models on the three datasets. Amongst the finite topic models, the Gaussian softmax construction (GSM)

[2] http://labrosa.ee.columbia.edu/millionsong/musixmatch (Bertin-Mahieux et al., 2011)

[3] http://qwone.com/ jason/20Newsgroups

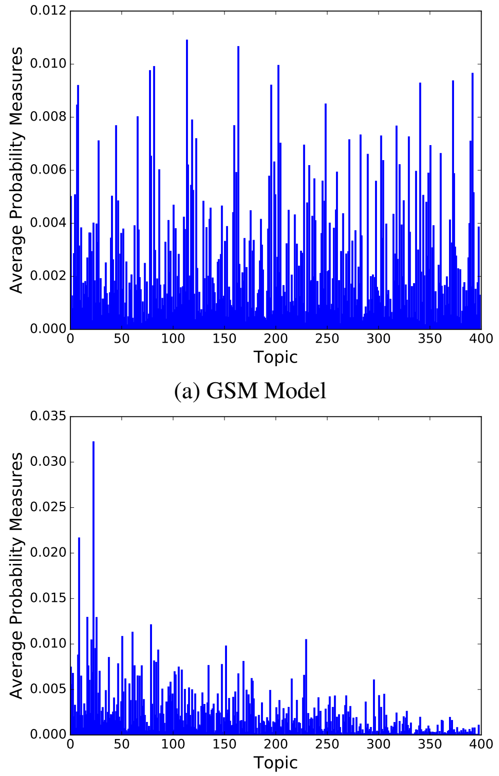
[4] http://trec.nist.gov/data/reuters/reuters.html

[5] We use the vocabulary provided by Srivastava & Sutton (2016) for direct comparison.

achieves the lowest perplexity in most cases, while all of the GSM, GSB and RSB models are significantly better than the benchmark LDA and NVLDA models. Amongst our selection of unbounded topic models, we compare our truncation-free RSB model, which applies an RNN to dynamically increase the active topics (γ is empirically set as $5e^{-5}$), with the traditional non-parametric HDP topic model (Teh et al., 2006). Here we see that the recurrent neural topic model performs significantly better than the HDP topic model on perplexity.

Next we evaluate our neural network parameterisations as document models with the implicit topic distribution introduced in Section 3.3. Table 2 compares the proposed neural document models with the benchmarks. According to our experimental results, the generalisation abilities of the GSM, GSB and RSB models are all improved by switching to an implicit topic distribution, and their performance is also significantly better than the NVDM and ProdLDA. We hypothesise that this effect is due to the models not needing to infer the topic-word assignments, which makes optimisation much easier. Interestingly, the RSB model performs better than the GSM and GSB on *20NewsGroups* in both the 50 and 200 topic settings. This is possibly due to the fact that GSM and GSB apply linear transformations W_1 and W_2 to generate the hidden variable θ and breaking proportions η from a Gaussian draw, while the RSB applies recurrent neural networks to produce η in a sequence which induces dependencies in η and helps escape local minima. It is worth noting that the recurrent neural network uses more parameters than the other two models. As mentioned in Section 3.3, GSM is a variant of NVDM that applies topic and word vectors to construct the topic distribution over words instead of directly modelling a multinomial distribution by a softmax function, which further simplifies optimisation. If it is not necessary to model the explicit topic distribution over words, using an implicit topic distribution may lead to better generalisation.

To further demonstrate the effectiveness of the stick-breaking construction, Figure 5 presents the average probability of each topic by estimating the posterior probability $q(z|d)$ of each document from *20NewsGroups*. Here we set the number of topics to 400, which is large enough for this dataset. Figure 5a shows that the topics with higher probability are evenly distributed. While in Figure 5a the higher probability ones are placed in the front, and we can see a small tail on the topics after 300. Due to the sparsity inducing property of the stick-breaking construction, the topics on the tail are less likely to be sampled. This is also the advantage of stick-breaking construction when we apply the RSB-TF as a non-parameteric topic model, since the model activates the topics according to the knowledge learned from data and it becomes less sensitive to the hyperparameter controlling the initial number of topics. Figure 6 shows the impact on test perplexity for the neural topic models when the maximum number of topics is increased. We can see that the performance of the GSM model gets worse if the maximum number of topics exceeds 400, but the GSB and RSB are stable even though the number of topics far outstrips that which the model requires. In addition, the RSB model performs better than GSB when the number of topics is under 200, but it becomes slightly worse than GSB when the number exceeds 400, possibly due to the difficulty of learning long sequences with RNNs.



(a) GSM Model

(b) GSB Model

Figure 5. Corpus level topic probability distributions.

Figure 7 shows the convergence process of the truncation-free RSB (RSB-TF) model on the *20NewsGroups*. With different initial number of topics, 10, 30, and 50. The RSB-TF dynamically increases the number of active topics to achieve a better variational lower bound. We can see the training perplexity keeps decreasing while the RSB-TF activates more topics. The numbers of active topics will stabilise when the convergence point is approaching (normally between 200 and 300 active topics on the *20NewsGroups*). Hence, as a non-parametric model, RSB-TF is not sensitive to the initial number of active topics.

In addition since the quality of the discovered topics is not directly reflected by perplexity (i.e. a function of log-likelihood), we evaluate the **topic observed coherence** by normalised point-wise mutual information (NPMI) (Lau

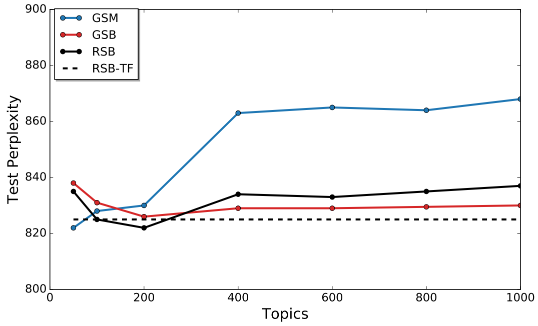


Figure 6. Test perplexities of the neural topic models with a varying maximum number of topics on the 20NewsGroups dataset. The truncation-free RSB (RSB-TF) dynamically increases the active topics, we use a dashed line to represent its test perplexity for reference in the figure.

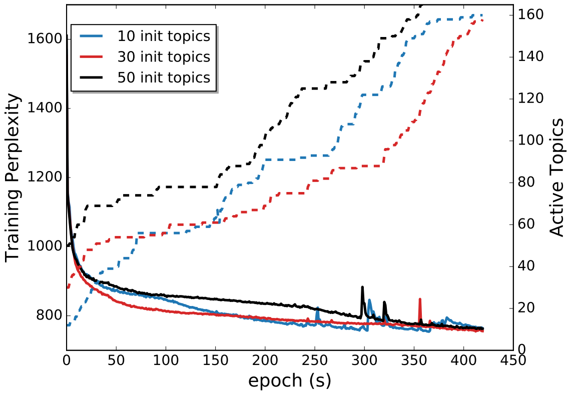


Figure 7. The convergence behavior of the truncation-free RSB model (RSB-TF) with different initial active topics on *20NewsGroups*. Dash lines represent the corresponding active topics.

et al., 2014). Table 3 shows the topic observed coherence achieved by the finite neural topic models. According to these results, there does not appear to be a significant difference in topic coherence amongst the neural topic models. We observe that in both the GSB and RSB, the NPMI scores of the former topics in the stick breaking order are higher than the latter ones. It is plausible as the stick-breaking construction implicitly assumes the order of the topics, the former topics obtain more sufficient gradients to update the topic distributions. Likewise we present the results obtained by the neural document models with implicit topic distributions. Though the topic probability distribution over words does not exist, we could rank the words by the positiveness of the connections between the words and each dimension of the latent variable. Interestingly the performance of these document models are significantly better than their topic model counterparts on topic coherence. The results of RSB-TF and HDP are not presented due to the fact that the number of active topics is dynamic, which makes these two models not directly comparable to the others. To further demonstrate the quality of the topics, we produce a t-SNE projection for the estimated topic proportions of each document in Figure 8.

Topic Model	Topics	
	50	200
GSM	0.121	0.110
GSB	0.095	0.081
RSB	0.111	0.097
OnlineLDA	0.131	0.112
NVLDA	0.110	0.110
Document Model	Latent Dimension	
	50	200
GSM	0.223	0.186
GSB	0.217	0.171
RSB	0.224	0.177
NVDM	0.186	0.157
ProdLDA	0.240	0.190 [6]

Table 3. Topic coherence on *20NewsGroups* (higher is better). We compute coherence over the top-5 words and top-10 words for all topics and then take the mean of both values.

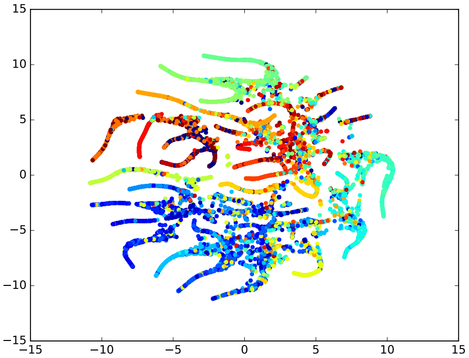


Figure 8. t-SNE projection of the estimated topic proportions of each document (i.e. $q(\theta|d)$) from *20NewsGroups*. The vectors are learned by the GSM model with 50 topics and each color represents one group from the 20 different groups of the dataset.

6. Conclusion

In this paper we have introduced a family of neural topic models using the Gaussian Softmax, Gaussian Stick-Breaking and Recurrent Stick-Breaking constructions for parameterising the latent multinomial topic distributions of each document. With the help of the stick-breaking construction, we are able to build neural topic models which exhibit similar sparse topic distributions as found with traditional Dirichlet-Multinomial models. By exploiting the ability of recurrent neural networks to model sequences of unbounded length, we further present a truncation-free variational inference method that allows the number of topics to dynamically increase. The evaluation results show that our neural models achieve state-of-the-art performance on a range of standard document corpora.

[6] The best scores we obtained are 0.222 and 0.175 for 50 and 200 topics respectively, but here we report the higher scores from Srivastava & Sutton (2016).

References

Attias, Hagai. A variational bayesian framework for graphical models. In *Proceedings of NIPS*, 2000.

Beal, Matthew James. *Variational algorithms for approximate Bayesian inference*. University of London, 2003.

Bertin-Mahieux, Thierry, Ellis, Daniel P.W., Whitman, Brian, and Lamere, Paul. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

Blei, David M and Lafferty, John D. Dynamic topic models. In *Proceedings of ICML*, pp. 113–120. ACM, 2006.

Blei, David M and Lafferty, John D. A correlated topic model of science. *The Annals of Applied Statistics*, 2007.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Bryant, Michael and Sudderth, Erik B. Truly nonparametric online variational inference for hierarchical dirichlet processes. In *Proceedings of NIPS*, 2012.

Carlin, Bradley P and Polson, Nicholas G. Inference for nonconjugate bayesian models using the gibbs sampler. *Canadian Journal of statistics*, 19(4):399–405, 1991.

Dieng, Adji B, Wang, Chong, Gao, Jianfeng, and Paisley, John. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*, 2016.

Hinton, Geoffrey E and Salakhutdinov, Ruslan. Replicated softmax: an undirected topic model. In *Proceedings of NIPS*, 2009.

Hoffman, Matthew, Bach, Francis R, and Blei, David M. Online learning for latent dirichlet allocation. In *Proceedings of NIPS*, pp. 856–864, 2010.

Hofmann, Thomas. Probabilistic latent semantic indexing. In *Proceedings of SIGIR*, 1999.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Khan, Mohammad Emtiyaz, Mohamed, Shakir, Marlin, Benjamin M, and Murphy, Kevin P. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *Proceedings of AISTATS*, 2012.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. In *Proceedings of ICLR*, 2014.

Landauer, Thomas K, Foltz, Peter W, and Laham, Darrell. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

Larochelle, Hugo and Lauly, Stanislas. A neural autoregressive topic model. In *Proceedings of NIPS*, 2012.

Lau, Jey Han, Newman, David, and Baldwin, Timothy. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of EACL*, pp. 530–539, 2014.

LeCun, Yann, Chopra, Sumit, and Hadsell, Raia. A tutorial on energy-based learning. *Predicting structured data*, 2006.

Mcauliffe, Jon D and Blei, David M. Supervised topic models. In *Advances in neural information processing systems*, pp. 121–128, 2008.

Miao, Yishu, Yu, Lei, and Blunsom, Phil. Neural variational inference for text processing. In *Proceedings of ICML*, 2016.

Mnih, Andriy and Gregor, Karol. Neural variational inference and learning in belief networks. In *Proceedings of ICML*, 2014.

Mnih, Volodymyr, Heess, Nicolas, and Graves, Alex. Recurrent models of visual attention. In *Proceedings of NIPS*, 2014.

Nalisnick, Eric and Smyth, Padhraic. Deep generative models with stick-breaking priors. *arXiv preprint arXiv:1605.06197*, 2016.

Rezende, Danilo J, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of ICML*, 2014.

Rosen-Zvi, Michal, Griffiths, Thomas, Steyvers, Mark, and Smyth, Padhraic. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494. AUAI Press, 2004.

Sethuraman, Jayaram. A constructive definition of dirichlet priors. *Statistica sinica*, pp. 639–650, 1994.

Srivastava, Akash and Sutton, Charles. Neural variational inference for topic models. *Bayesian deep learning workshop, NIPS 2016*, 2016.

Teh, Yee Whye, Jordan, Michael I, Beal, Matthew J, and Blei, David M. Hierarchical dirichlet processes. *Journal of the American Statistical Asociation*, 101(476), 2006.

Wang, Chong and Blei, David M. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031, 2013.

Wang, Chong, Paisley, John William, and Blei, David M. Online variational inference for the hierarchical dirichlet process. In *Proceedings of AISTATS*, 2011.

Wang, Xuerui and McCallum, Andrew. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433. ACM, 2006.

Xie, Pengtao, Deng, Yuntian, and Xing, Eric. Diversifying restricted boltzmann machine for document modeling. In *Proceedings of KDD*, pp. 1315–1324. ACM, 2015.

A. Discovered Topics

Table 4 presents the topics by the words with highest probability (top-10 words) achieved by different neural topic models on *20NewsGroups* dataset.

Space	***Religion***	***Encryption***	***Sport***	***Science***
space	god	encryption	player	science
satellite	atheism	device	hall	theory
april	exist	technology	defensive	scientific
sequence	atheist	protect	team	universe
launch	moral	americans	average	experiment
president	existence	chip	career	observation
station	marriage	use	league	evidence
radar	system	privacy	play	exist
training	parent	industry	bob	god
committee	murder	enforcement	year	mistake

(a) Topics learned by GSM.

Space	***Religion***	***Lawsuit***	***Vehicle***	***Science***
moon	atheist	homicide	bike	theory
lunar	life	gun	motorcycle	science
orbit	eternal	rate	dod	gary
spacecraft	christianity	handgun	insurance	scientific
billion	hell	crime	bmw	sun
launch	god	firearm	ride	orbit
space	christian	weapon	dealer	energy
hockey	atheism	knife	oo	experiment
cost	religion	study	car	mechanism
nasa	brian	death	buy	star

(b) Topics learned by GSB.

Aerospace	***Crime***	***Hardware***	***Technology***	***Science***
instruction	gun	drive	technology	science
spacecraft	weapon	scsi	americans	hell
amp	crime	ide	pit	scientific
pat	firearm	scsus	encryption	evidence
wing	criminal	hd	policy	physical
plane	use	go	industry	eternal
algorithm	control	controller	protect	universe
db	handgun	tape	privacy	experiment
reduce	law	datum	product	reason
orbit	kill	isa	approach	death

(c) Topics learned by RSB.

Table 4. Topics learned by neural topic models on 20NewsGroups dataset.

B. Topic Diversity

An issue that exists in both probabilistic and neural topic models is redundant topics. In neural models, however, we are able to straightforwardly regularise the distance between each of the topic vectors in order to diversify the topics. Following Xie et al. (2015), we apply such topic diversity regularisation during the inference process. We compute the angles between each two topics $a(t_i, t_j) = arccos(\frac{|t_i \cdot t_j|}{||t_i|| \cdot ||t_j||})$. Then, the mean angle of all pairs of K topics is $\zeta = \frac{1}{K^2}\sum_i \sum_j a(t_i, t_j)$, and the variance is $\nu = \frac{1}{K^2}\sum_i \sum_j (a(t_i, t_j) - \zeta)^2$. We add the following topic diversity regularisation to the variational objective:

$$\mathcal{J} = \mathcal{L} + \lambda(\zeta - \nu),$$

where λ is a hyper-parameter for the regularisation that is set as 0.1 in the experiments. During training, the mean angle is encouraged to be larger while the variance is suppressed to be smaller so that all of the topics will be pushed away from each other in the topic semantic space. Though in practice diversity regularisation does not provide a significant improvement to perplexity (2∼5 in most cases), it helps reduce topic redundancy and can be easily applied on topic vectors instead of the simplex over the full vocabulary.